Master of Information and Data Science
DATASCI 281: Computer Vision
Spring 2022

Professor Hany Farid
University of California, Berkeley

Expectation Maximization

The expectation/maximization (EM) algorithm simultaneously groups and fits data generated from multiple parametric models. Consider the data points in Figure 1(a). You can clearly see that these data points are well fit to one of two possible lines (our models). Specifically, each data point, $(x(i), y(i))$ with $i \in [1, n]$, is generated from one of two models given by the equation of a line:

$$y(i) = a_1 x(i) + b_1 + n_1(i) \quad (1)$$
$$y(i) = a_2 x(i) + b_2 + n_2(i), \quad (2)$$

where the model parameters are a_1, b_1 and a_2, b_2. We assume some amount of imperfection in the data or underlying model, which we approximate with additive noise terms $n_1(i)$ and $n_2(i)$.

If we are given the model parameters (a_1, b_1 and a_2, b_2), then determining which data point i was generated by which model would be a simple matter of choosing the model k that minimizes the error between the data and the model prediction:

$$r_k(i) = |a_k x(i) + b_k - y(i))|, \quad (3)$$

for $k = 1, 2$ in our example. That is, we simply ask, to which line is each data point closest, Figure 1(b). We measure this error using the vertical distance between the data point and the model, as opposed to the orthogonal distance.

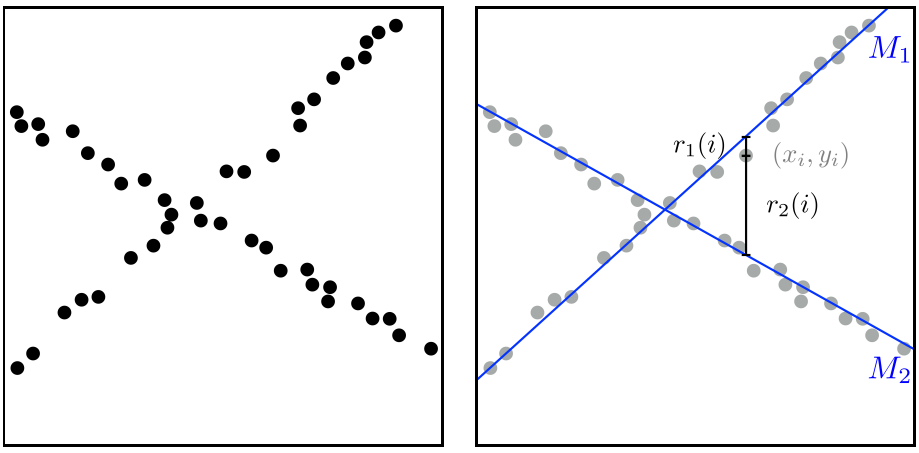

Figure 1: Fit two lines to the data points. The data point (x_i, y_i) has a residual error $r_1(i)$ and $r_2(i)$ for each model (line) M_1 and M_2.

This is because this error lends itself to a least-squares estimation, described next.

On the other hand, if we are given which data points were generated by which model, then estimating the model parameters reduces to solving, for each model k, an over-constrained set of linear equations:

$$\begin{pmatrix} x_k(1) & 1 \\ x_k(2) & 1 \\ \vdots & \vdots \\ x_k(n) & 1 \end{pmatrix} \begin{pmatrix} a_k \\ b_k \end{pmatrix} = \begin{pmatrix} y_k(1) \\ y_k(2) \\ \vdots \\ y_k(n) \end{pmatrix}, \tag{4}$$

where the $x_k(i)$ and $y_k(i)$ all belong to model k.

In either case, knowing one piece of information (the model assignment or parameters) makes determining the other relatively easy. But, lacking either piece of information makes this a considerably more difficult estimation problem. The EM algorithm is an iterative two step algorithm that estimates both the model assignment and model parameters.

The E-step of EM assumes that the model parameters are known (initially, the parameters can be assigned random values) and calculates the probability of each data point belonging to each model. In so doing the model assignment is made in a probabilistic fashion. Each data point is not explicitly assigned ownership to a single model, rather each data point i is assigned a probability of membership in model k. For each model the residual

error is first computed as:

$$r_k(i) \;=\; a_k x(i) + b_k - y(i) \tag{5}$$

from which the probabilities are calculated. We ask, what is the probability of point i belonging to model k given the residual error. For our two model example:[1]

$$P(a_k, b_k \mid r_k(i)) \;=\; \frac{P(r_k(i) \mid a_k, b_k)}{P(r_1(i) \mid a_1, b_1) + P(r_2(i) \mid a_2, b_2)}, \tag{6}$$

for $k = 1, 2$. If we assume a Gaussian probability distribution for the noise $n_k(i)$, then the probability takes the form:

$$w_k(i) = P(a_k, b_k | r_k(i)) = \frac{e^{-r_k^2(i)/2\sigma^2}}{e^{-r_1^2(i)/2\sigma^2} + e^{-r_2^2(i)/2\sigma^2}}, \tag{7}$$

where, σ is proportional to the amount of noise in the data.

The M-step of EM takes the probability of each data point belonging to each model and re-estimates the model parameters using weighted least-squares. The following weighted quadratic error function on the model parameters is minimized:

$$E(a_k, b_k) \;=\; \sum_{i=1}^{n} \left(w_k(i)(a_k x(i) + b_k - y(i))\right)^2. \tag{8}$$

The intuition here is that each data point contributes to the estimation of each model's parameters in proportion to the belief in its membership in that particular model.

This quadratic error function can be rewritten in matrix form:

$$E(\vec{m}_k) \;=\; \|W_k(X\vec{m}_k - \vec{y})\|^2, \tag{9}$$

where the model parameters are $\vec{m}_k = \begin{pmatrix} a_k & b_k \end{pmatrix}$, the $n \times 2$ matrix X is:

$$X \;=\; \begin{pmatrix} x(1) & 1 \\ x(2) & 1 \\ \vdots & \vdots \\ x(n) & 1 \end{pmatrix}, \tag{10}$$

[1]The expression $P(a_k, b_k \mid r_k(i))$ is the conditional probability of *observing* the model parameters a_k, b_k *given* the residual error $r_k(i)$ for each data point i. If, for example, the residual error is zero, then the probability will be high. As the residual error increases, the probability decreases proportionally because the sample does not satisfy the model. The expansion of the conditional probability is from Bayes' rule: $P(X \mid Y_k) = \frac{P(Y_k \mid X)P(X)}{\sum_l P(Y_l \mid X)P(X)}$.

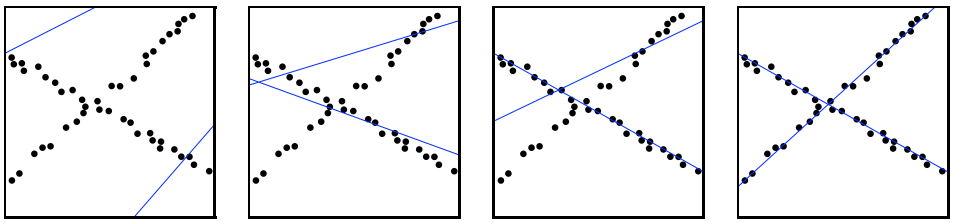

Figure 2: Four iterations of the EM algorithm.

the $n \times 1$ vector y is:

$$\vec{y} \quad = \quad \begin{pmatrix} y(1) & y(2) & \dots & y(n) \end{pmatrix}^T, \tag{11}$$

and W_k is a $n \times n$ diagonal weighting matrix:

$$W_k \quad = \quad \begin{pmatrix} w_k(1) & 0 & 0 & \dots & 0 \\ 0 & w_k(2) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & w_k(n) \end{pmatrix} \tag{12}$$

The quadratic error function is minimized by computing the gradient with respect to the model parameters, setting the result equal to zero and solving for the model parameters. This yields the weighted least squares solution:

$$\vec{m}_k \quad = \quad (X^T W_k^T W_k X)^{-1} X^T W_k^T W_k \vec{y}, \tag{13}$$

The EM algorithm iteratively executes the E- and M-step, repeatedly estimating and refining the model assignments and parameters. Several iterations of EM applied to fitting data generated from two linear models are shown in Figure 2. The model parameters are initially assigned random values, and after a few iterations the algorithm converges to a solution.

The EM algorithm is guaranteed to converge. The convergence to the right or desired solution, however, depends on the quality of the initial model parameters used to bootstrap the EM iterations.

The EM algorithm can be sensitive to the value of σ used, Equation (7). It is recommended that with a reasonable starting value, the value of σ_k can be updated on each EM iteration as:

$$\sigma_k \quad = \quad \frac{\sum_{i=1}^{n} w_k(i) r_k^2(i)}{\sum_{i=1}^{n} w_k(i)}. \tag{14}$$

If the value of σ is initially too small, then it is unlikely that data points will be assigned to the correct model. On the other hand, if σ is initially too large, then data points will be equally likely to belong to either model. In either case, the convergence of EM to the desired solution will be hampered by an inappropriate value of σ.